VALIDITIES AND RELIABILITIES OF SELECTED ACHIEVEMENT SKILLS TESTS IN SOFTBALL FOR USE AT THE UNIVERSITY OF MINNESOTA

A Field Report

Presented to

The Graduate Division

Drake University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Education

---

by

Jacqueline Shick

August 1964

VALIDITIES AND RELIABILITIES OF SELECTED ACHIEVEMENT SKILLS TESTS IN SOFTBALL FOR USE AT THE UNIVERSITY OF MINNESOTA

by

Jacqueline Shick

Approved by Committee:

Chairman

Dean of the Graduate Division

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

The educational process is composed of many factors. Evaluation is one such factor and has a unique contribution to make to the development of the learner. Educators know that effective teaching must relate to the objectives of any given program. Therefore, physical educators must evaluate the social, emotional, mental, and physical development of the learner.

The latter facet is of prime importance, since the basic principles of the field are directed toward educating the individual through the physical. Evaluation of the physical must include not only the measurement of the growth and development but also, and perhaps even more important, measurements of motor skill for, ". . . the learning of desirable skills is the very foundation of physical education."[1]

Much evaluation of skill is done by observation. This observation may be incidental or it may be directed and systematic. However, a more objective means of evaluation is needed to supplement that of observation. This objective supplement takes the form of skills tests.

---

[1] Harrison Clarke, Application of Measurement to Health and Physical Education, (New York: Prentice-Hall, Inc., 1959), p. 324.

"Skills tests reflect the ability of the pupil to perform in a specified sport such as badminton, handball, or basketball."[1] As with written tests, good skills tests must be both valid and reliable. That is, they must measure what they purport to measure and they must do so consistently. Another requirement for a good skills test item is that it must measure an ability considered to be an essential element of the game for which the test is devised. Even if this has been established, there are still possibilities of error. A student may take a skills test and achieve a satisfactory score and yet perform poorly in the game situation. Physical educators do not feel that this would be extremely unusual due to the complexity and interaction of all operative variables. A student might exhibit considerable ability in any or several of the separate elements of a given sport but lack the ability to combine these skills into a unified whole when placed under the stress of the game situation.[2] For this reason, it is difficult to obtain high validities on skills test items. The converse of this would rarely be true. That is, students who perform well in a game situation should perform well in an adequately devised test.

---

[1]Donald K. Mathews, Measurement in Physical Education (Philadelphia: W.B. Saunders Company, 1958), p. 161.

[2]Carl Willgoose, Evaluation in Health Education and Physical Education, (Philadelphia: W.B. Saunders Co., 1959), p. 216.

Brownell and Hagman stated that even though available skills tests are not able to make fine discriminations concerning the player's ability in the game situation, they still do serve a useful purpose. These skills tests may point out skill weaknesses and serve to aid in the grouping of players in reference to levels of achievement.[1] Stroup supports this by saying that although, "A completely valid test for a sport is unlikely, sports skills tests can be very useful."[2]

Standardized skills tests have been devised for most of the sports taught in the schools. However, at times, local conditions will not permit the use of these standardized tests. At such times, Stroup felt that individual teachers should devise their own items and check them for validity and reliability.[3]

Statement of problem. The measurement of specific skills in softball presented a problem to the writer, who is an instructor of physical education at the University of Minnesota. Softball classes for women are part of her teaching assignments. When preparing for skill evaluation,

---

[1]Clifford Lee Brownell and E. Patricia Hagman. Physical Education--Foundations and Principles, (New York: McGraw-Hill Book Company, Inc., 1951), p. 366.

[2]Francis Stroup, Measurement in Physical Education, (New York: The Ronald Press Company, 1957), p. 129.

[3]Ibid., p. 130.

the writer found only a limited number of tests available for the measurement of softball skills. Those that have been devised are not practical for use in this situation due to an excessive time requirement or impractical demands for facilities. Thus, the problem became one of establishing validities and reliabilities for selected softball achievement tests which could be used at the University of Minnesota.

Importance of the study. Such tests would be of value since they could serve a number of purposes. They could be used in the following ways:

I. In college softball classes at the University of Minnesota.
   A. To measure achievement in softball skills.
   B. As motivational devices.
   C. To group players into teams in reference to level of achievement.

II. In methods courses for major students.
   A. As an evaluative technique to measure levels of achievement.
   B. As an instructional device which students might use later in their own teaching experiences.

III. As competency tests for major students. (Such tests are given to major students who feel they can show enough skill to warrant exemption from classes in specific areas.)

IV. At other colleges and universities with the same or similar facilities.

CHAPTER II

REVIEW OF LITERATURE

A review of the literature reveals a consensus that skills tests are an important part of the teaching procedure. Willgoose stated, "All great professions have achieved status through measurement. . . the use of valid and precise measuring instruments is an absolute prerequisite for gaining recognition as a profession."[1] The same author believed that all schools, large and small, should try to find the best tests available, to begin using them and improving them, and to evaluate the results.[2]

In spite of the fact that physical educators recognize the need for skills tests, a review of research reveals that there are few such tests designed to measure achievement in softball skills.

In the Official Softball-Volleyball Guide published for the National Section on Women's Athletics of the American Association for Health, Physical Education and Recreation in 1943, Jesseline Thomas reported on the following skills tests: repeated throws, throw for distance, field and throw to base, batting, baserunning, and

---

[1] Willgoose, op. cit., p. 44.

[2] Ibid., p. 45.

pitching. The first two tests (repeated throws and throw for distance) were also included in the book Evaluation in Physical Education by M. Gladys Scott and Esther French. A third test (field and throw to base) was also presented by both Scott and French and Thomas, but some differences do appear. Descriptions of the tests are as follows:

1. The repeated throws test is a timed test and the number of good throws completed within the time limit is recorded. The person being tested stands behind a restraining line drawn on the floor at a distance of fifteen feet from the wall and parallel to it. A line is drawn on the wall at a height of seven feet six inches from the floor. The starting signal is, "Ready? Go!" At the signal, the girl being tested throws the ball against the wall so that it hits above the seven foot six inch line, catches it, and repeats the action as many times as she can in the alloted thirty seconds. To be considered "good", the throw must be made from behind the restraining line, (to step on or over the line is considered a foot fault) and must hit the wall on or above the seven foot six inch line. If the girl loses control of the ball during the test, she must recover it herself. The only penalty is the loss

of time. Six trials of thirty seconds each are given. One point is given for each "good" throw and the score for the test is the total of six trials. Scott and French report a reliability coefficient of .89 for this test when the odd even method is used for college women. Application of the Spearman-Brown formula raises the coefficient to .94. These same authors report that lower coefficients were obtained when a similar test was given to junior high school girls. The validity coefficients obtained were .64 for the junior high school girls and .51 for the college women when a subjective rating was compared with scores on the repeated throws test.[1] This subjective rating ranged from "7-excellent, to - 1 - dub, given the players by their instructors."[2] Scott and French felt that although the repeated throws is highly reliable, it does not give a clear differentiation among students who fall in the middle ability groups.[3]

---

[1] Jesseline Thomas, "Skill Tests, (Official Softball-Volleyball Guide (New York: A.S. Barnes and Company, 1943), pp. 22-4; and M. Gladys Scott and Esther French, Evaluation in Physical Education, (St. Louis: The C.V. Mosby Company, 1950), pp. 112-4.

[2] Thomas, op. cit., p. 22.

[3] Scott and French, op. cit., p. 114.

2. The throw for distance test is, as the name implies, a test to see how far the ball can be thrown using an overhand or sidearm throw. The testee may take one step when making the throw but that step may not be over a specified line. One trial consists of three throws; however, only the best throw is measured and recorded. The subject is allowed three trials, which means she takes nine throws in all. "The throw is measured as the distance in feet from the starting line to the spot where the ball first touches the ground."[1] The score is the best of the three recorded throws. A reliability of .95 was obtained on successive trials using a population of seventh and eighth grade girls. Validity coefficients were computed using ratings and achievement scores on this test. A coefficient of .81 was found when seventh and eighth grade girls were the population but the validity dropped to .63 when college women were used as subjects. Research has shown that there is a relationship between the speed with which an object is thrown and the distance that it travels. That is, the faster the ball is thrown, the further it will

---

[1] Ibid., p. 114.

tend to travel. Because this relationship exists, and since speed of the throw is many times a vital factor in the game situation, the ability to throw long distances is an important skill to be measured. Scott and French feel that, "This appears to be the best single test to yet be devised for measuring softball playing ability."[1]

An intercorrelation between the aforementioned two tests (repeated throw and throw for distance) brought a correlation of .81 and so it is felt that only one of these tests need be given.

3. The diagrams for the fielding test are given in Figure 1. The reader will note that the differences between the diagrams involve the distance that the base is from the two walls and the dimensions and the number of the concentric circles of the target. The dimensions of the circles as described by Scott and French are radii of three inches, twelve inches, twenty-one inches, thirty inches and thirty-nine inches respectively. The dimensions of the circles as described by Thomas are diameters of six inches, twenty-two inches, forty-two inches, and

---

[1] Ibid., p. 115.

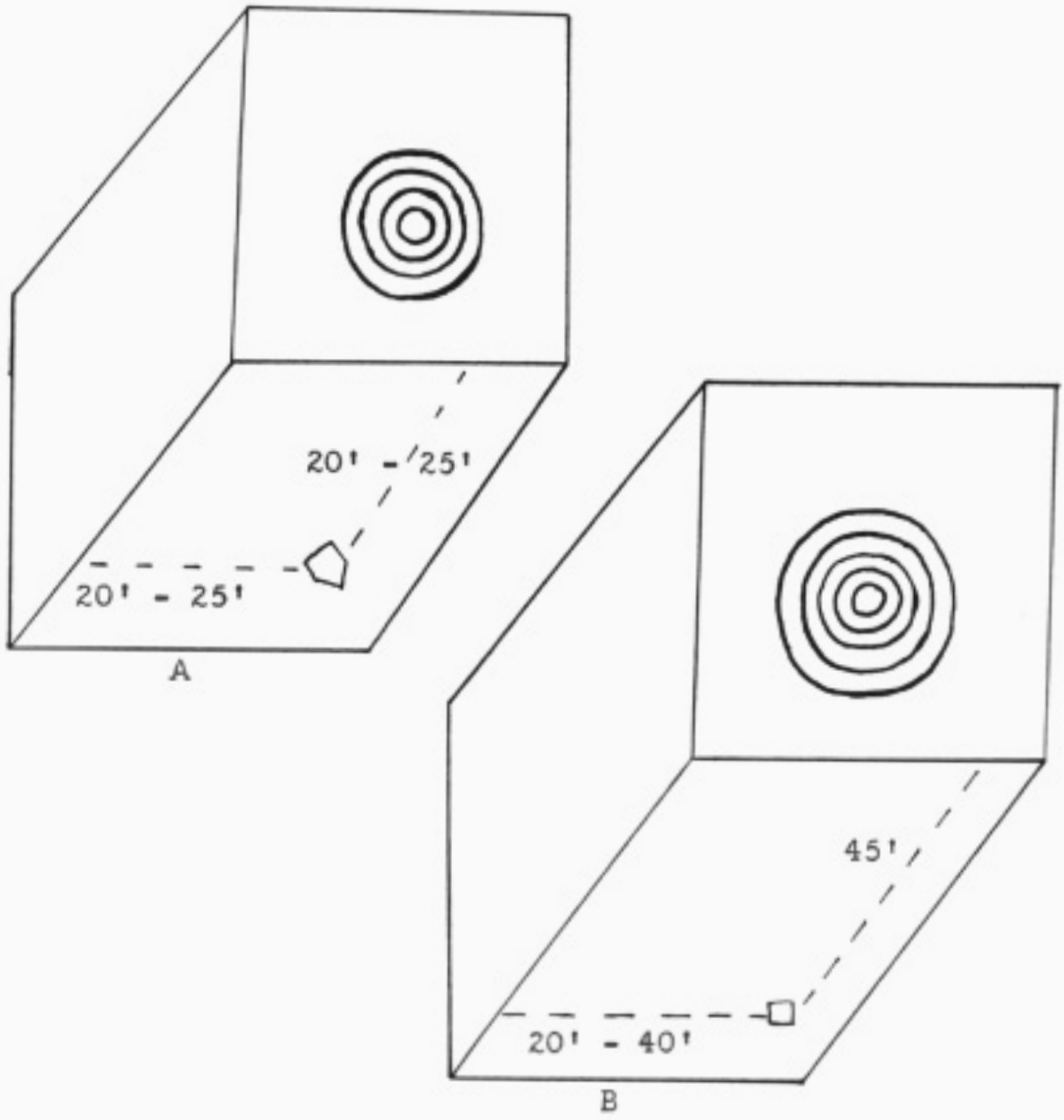


Figure 1. Dimensions and targets for fielding tests as described by Scott and French (B) and by Thomas (A).

sixty-six inches respectively. For both targets, the center of the target is to be forty-eight inches from the floor. The task which the testee must do is exactly the same when described by Scott and French and by Thomas. The test itself consists of ten trials. From a starting position on the base, each trial involves throwing the ball at the blank wall, fielding it, returning to touch the base, and then throwing the ball at the target. Each trial is timed from the starting signal until the moment when the ball hits the wall target. For this reason, the scoring of the test has two aspects, the total time for all ten trials and the total of the scores made on the ten trials. The scoring would differ on the second aspect of the two tests in that the one described by Scott and French would have a maximum total of fifty and for the one described by Thomas, a maximum total of forty. Thomas listed no validities or reliabilities but Scott and French did. With a population of freshmen and sophomore women, they found a reliability of .56 using the odd-even method for computation and using the Spearman-Brown formula for correction. This same population was rated on throwing form and footwork.

When correlated, these ratings produced a validity of .49 with the test.[1]

4. The fourth test described was a batting test. It consists of ten trials to hit a legally pitched ball. An umpire is used to call "balls" and "strikes." "Balls" are disregarded, but the "strikes" count as a trial. The same pitcher and umpire are used for the entire testing period. The testee's score is the sum of the ten trials. The scoring is as follows: an outfield hit equals five points, an infield hit equals three points, a foul equals one point, and a strike (either a swinging strike or a called strike) equals zero. With a population of one hundred eighteen junior high school girls, a reliability of .65 was found, but by using the Spearman-Brown correction the coefficient became .79. Using the subjective ratings of teacher and students, a validity coefficient of .72 was obtained.[2]

5. In the baserunning test, the subject stands in the batter's box, holding a bat. When the timer says, "Go!" she must swing the bat and

---

[1] Ibid., pp. 115-118; and Thomas, op. cit., pp. 24-25.

[2] Ibid., p. 25.

then run all of the bases in order. Two trials are administered. The better of the two trials is used. For actual scoring purposes, the timing is done from the signal, "Go!" until the moment the runner's foot touches second base. The time is recorded to the nearest tenth of a second. Using the one hundred eighteen junior high girls as subjects, the reliability and validity coefficients were .71 and .55 respectively. Student and teacher subjective ratings were used to determine validity.[1]

6. The pitching test used an organization shown in Figure 2. Groups I and II constitute one squad. The test is administered in the following manner:
   a. The pitcher in Group I (labeled catcher) pitches to the catcher in Group II. The umpire of Group II calls the pitch and the recorder in Group II records the call.
   b. The procedure is reversed and the catcher from Group II becomes the pitcher, while the pitcher from Group I becomes the catcher for that group.

   When the two subjects have each completed thirty pitches, the girls in each group rotate one

---

[1] Ibid., pp. 25-6.

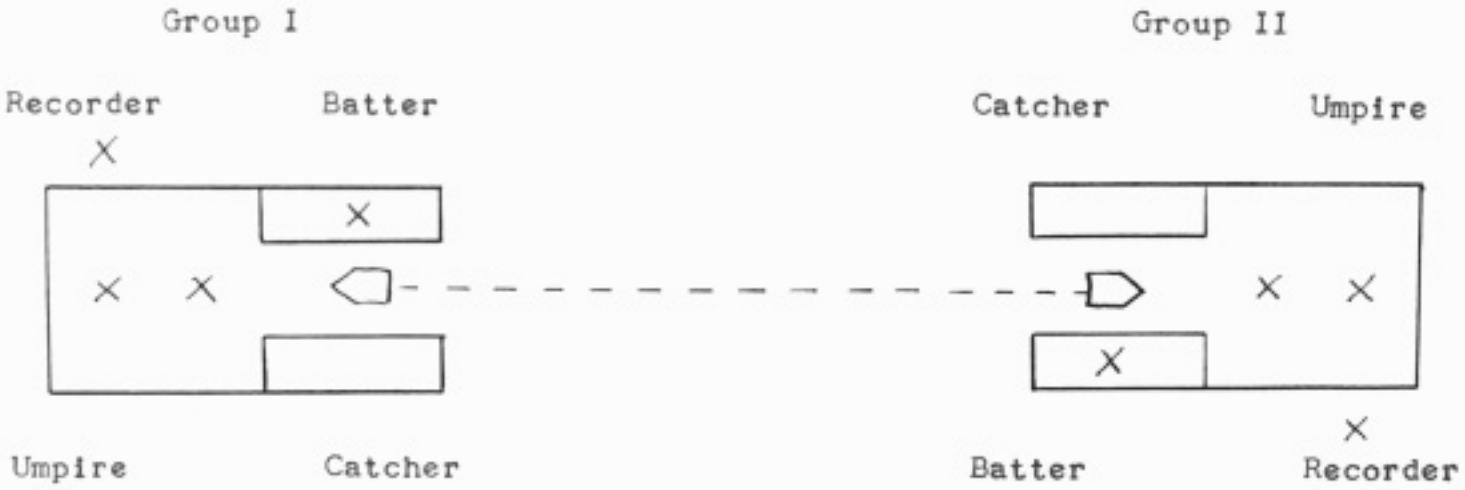


Figure 2. Organization for pitching test as described by Thomas.

position in a counterclockwise manner. Each testee is allowed three practice throws. No reliabilities or validities for this test were reported.[1]

Other research related to softball skills tests was reported in the December, 1931, issue of the Research Quarterly. Elizabeth Rodgers and Marjorie Heath reported on "An Experiment in the Use of Knowledge and Skill Tests in Playground Baseball." The writers developed a battery of skill tests which include the following items:

1. Baseball throw for accuracy (underhand throw using a twelve inch playground baseball).
2. Batting with an actual pitcher.
3. Catching fly balls (repeated throws against a wall).
4. Catching grounders.
5. Hit and run to first base.

The population for these tests were boys in the fifth and sixth grades. Reliabilities were established by correlating scores on a pre-season test with those made on a test at the close of the season. Although no single test item gave a correlation of better than .69, a battery reliability of .83 was obtained when the testors used a formula "for correlation between scores of a series of

---

[1] Ibid., pp. 26-7.

tests not strictly comparable."[1] To establish validity, a combination of ratings by the teacher and squad leader as well as success in making class teams and school teams was used. The correlation for boys in grade five was .63 and for boys in grade six, .65.[2]

In a test of softball batting ability developed by Margaret Fox and Olive Young, the following three components were tested:

1. Batting for distance.
2. Placement of the ball.
3. Batting and running to first base.

A batting tee was used for all three tests. Three warm-ups were allowed for each test and five trials were then administered. A great deal of space is required to administer these tests, especially the batting for distance test. Reliability coefficients ranged from .81 to .96. The split halves method of computation was used and the Spearman-Brown formula was applied. To obtain the validity correlations, the test scores were correlated with the judges' ratings of batting ability. The judges rated the testees on grip, stance, swing, and consistency of hitting. The only one

---

[1]Elizabeth Rodgers and Marjorie Heath, "An Experiment in the Use of Knowledge and Skill Tests in Playground Baseball," Research Quarterly, II (December, 1931), 123.

[2]Ibid., pp. 113-125.

felt to be at all valid was the batting for distance test which had a coefficient of .64.[1]

Marion Broer administered selected skills tests in various sports and tests of motor ability to junior high school girls to determine the reliabilities of these tests. One of the sports selected was softball. Three softball skills tests were administered. They were:

1. Overhand throw for distance.
2. Underhand throw for accuracy.
3. Batting, using a tee.

In the overhand throw for distance, Broer correlated the first and second throws and found a correlation of .94. With this correlation, she states that only one trial is needed for this test. A circular target, forty-eight inches in diameter was used for the underhand throw for accuracy. This target consisted of three concentric circles. The testee throws from a distance of thirty feet and is allowed two trials of ten throws each. To determine reliability, odd and even scores were correlated. A coefficient of .42 was obtained. The batting test was administered using a tee. Trials one and three were correlated with trials two and four with the resulting correlation of .61.[2]

---

[1]Margaret Fox and Olive Young, "A Test of Softball Batting Ability," Research Quarterly, XXV (March, 1954), 26-7.

[2]Marian Broer, "Reliability of Certain Skills Tests for Junior High School Girls," Research Quarterly, XXIX (May, 1958), 139-45.

Marie Liba administered tests to determine if there is a systematic or random variation in trial to trial performance in tests of velocity in an overhand softball throw. By recording the distance and the time of the ball flight, she was able to compute the velocity of the throw. The test consisted of two warm-ups followed by five trials. She found no evidence of systematic variation from trial to trial.[1]

A softball throw for distance is included in the Barrow Motor Ability Test. The testing area is marked off at five-yard intervals. The testee takes three throws and the score of the three trials is recorded to the nearest foot.[2]

A similar test (softball throw for distance) was included in the American Association of Health, Physical Education, and Recreation Youth Fitness Test.[3]

C. H. McCloy included a softball test under tests of special abilities. A test of accuracy in throws at a target consisting of concentric circles is mentioned. No

---

[1]Marie Liba, "A Trend Test as a Preliminary to Reliability Estimation" Research Quarterly, XXIII (May, 1962), 245-8.

[2]Harold M. Barrow, "Test of Motor Ability for College Men," Research Quarterly, XXV (October, 1954), 253.

[3]AAHPER Youth Fitness Test Manual, American Association for Health, Physical Education, and Recreation, 1201 Sixteenth Street, N.W., Washington 6, D.C.

dimensions for the target face or distances from the target are given. He states that these will vary with the age and size of the group. McCloy stated that at least twenty-five trials are given. In his opinion, no study which has been done is considered adequate.[1]

Arthur J. Wendler administered tests using a softball to determine relationships to general motor educability and eye-hand co-ordination. Tests included in the study were fielding ground balls, throwing a softball, catching a fly ball, and target throwing. He found low, positive correlations between each of the tests and general motor educability and eye-hand co-ordination.[2]

The foregoing review of literature indicates that there is a need for the development of achievement tests in softball for college women. Thus, a study of this nature would be of value.

---

[1] C.H. McCloy, Tests and Measurements in Health and Physical Education, (New York: F.S. Crofts and Company, 1942), p. 155.

[2] C.H. McCloy and Norma D. Young, Tests and Measurements in Health and Physical Education, (New York: Appleton-Century-Crofts, Inc., 1954), pp. 101-2.

CHAPTER III

PROCEDURES

Selection of sample. The subjects used in this study were members of a non-professional class in softball. Since softball is not a required course, the class was composed of students who had elected to take it. These students were freshmen, sophomore, junior, and senior college women.

A total of thirty-one students registered for the softball class. However, the statistical computations do not use an N of 31 for each test because students who were absent were not included. This was done so that the testing conditions could be held constant.

This course taught by the writer was the only softball class offered during the quarter; there were no separate courses designed for beginning, intermediate and advanced students. Consequently, there was a wide range of achievement in the basic skills at the beginning of the quarter.

Sources of data. The writer devised four achievement tests which were administered. The tests were a repeated throws test, an overhand or sidearm target test for accuracy and force, a test for fielding ground balls, and

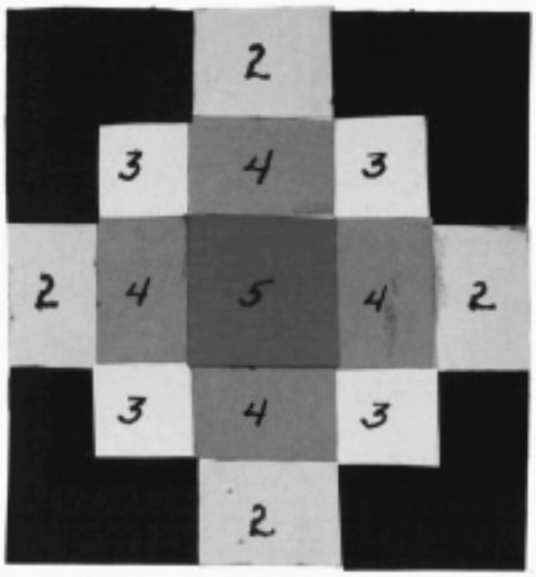
a timed base running test. A regulation softball was used for the first three tests.

The test for repeated throws was a timed test with a duration of thirty seconds. The testee stood behind a line drawn on the floor twenty-three feet from the wall and parallel to it. The starting signal was, "Ready? Set! Go!" At the signal, the girl being tested threw the ball at the wall, attempting to hit above a line ten feet from the floor. As the ball rebounded from the wall, the testee attempted to recover it either by catching it in the air or by fielding it from the floor. This procedure was repeated as many times as possible within the specified time limit. Each throw had to be made from behind the twenty-three-foot line. No point was given if the ball hit below the line on the wall or if the testee stepped on or over the line on the floor when throwing the ball. If the testee failed to field the ball, she was required to recover it herself, the only penalty being the loss of time. Each testee was given one practice throw and then timed for thirty seconds. The score which was recorded was the total number of hits minus the total number of errors.

The fielding test was similar to the repeated throws test in many respects. It was a timed event with the duration of thirty seconds. In this test, the testee

stood behind a line drawn on the floor fifteen feet from the wall and parallel to it. The starting signal was, "Ready? Set! Go!" At the signal the girl being tested threw the ball at the wall attempting to hit below a line four feet from the floor. As the ball rebounded from the wall, the testee attempted to field it. The above procedure was repeated as many times as possible within the thirty second time limit. Each throw had to be made from behind the fifteen-foot line. No point was given if the ball hit above the line on the wall or if the testee stepped on or over the line on the floor when throwing the ball. If the testee failed to field the ball, she was required to recover it herself, the only penalty being the loss of time. Each testee was given one practice throw and then timed for thirty seconds. The score which was recorded was the total number of hits minus the total number of errors.

The overhand or sidearm target test for accuracy and force consisted of ten throws at a sixty-six inch square wall target. The center of the target was thirty-six inches from the floor. The scoring areas and dimensions of the wall target and the target on the floor are shown in Figure 3. The testee stood behind a line drawn on the floor forty feet from the wall. She was allowed two practice throws and then took ten test throws. If the testee stepped on or over the line on the floor, she was



5 = 18" x 18"

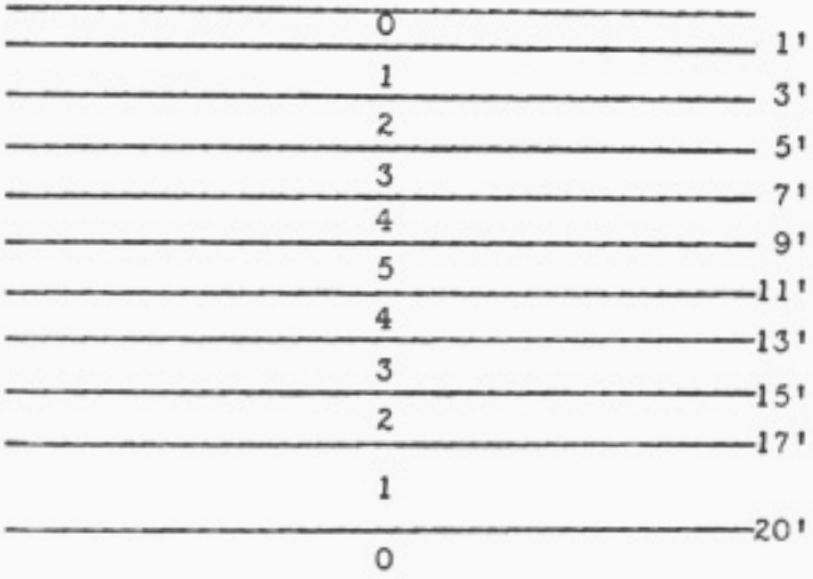
4 = 18" x 12"

3 = 12" x 12"

2 = 18" x 12"

1 = 24" x 24" (outside)

12"x12"x12"x12" (inside)



40'

Figure 3. Dimensions and target areas for the target test.

given a score of zero for the throw. Each throw was given two scores, a score for the hit on the wall and a score for the hit on the floor. Any hit outside the scoring areas was recorded as zero. The highest possible score for the wall target was fifty and the highest possible score for the floor target was fifty, making a possible sum total of one hundred for the ten trial test.

The baserunning test was a timed event administered on a regulation softball diamond. The testee stood on first base with one foot on the bag and the other off the bag and on the second-base side. At the sound of a whistle, blown by the writer, the girl started toward second base, touched it in passing, and ran to third base. The watch was started with the whistle and stopped as the testee first touched third base. Scores were recorded to the nearest tenth of a second.

Administration of tests. All tests were administered at the University of Minnesota. The repeated throws test, the fielding test, and the target test were administered in the fieldhouse of Norris Gymnasium, and the baserunning test was given outdoors. The first three tests were given before the student had had any opportunity to go outside for instructional classes. The baserunning test was given when the class met outside.

During the first meeting of the class, the writer told the students about the tests which she planned to administer and gave a short explanation of validity and reliability. At this time it was pointed out that practice affects statistical results and the students were instructed not to practice. The writer believes that the students complied with this request and did no practicing.

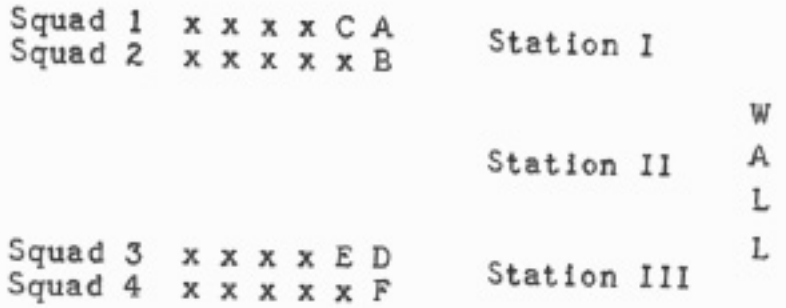
The second and third class periods were spent in instruction in throwing followed by "warming-up" by slowly increasing the distance and force of the throws.

The testing began on the fourth day of class. The writer attempted to administer the tests during the first part of the class period so that the rest of the period could be spent on other class work.

During the first testing period, the target test and the fielding test were given. The students were divided into four squads. Two squads took the fielding test and then rotated to the target test, while the other two squads took the tests in the reverse order. During the second testing period, the same two tests were administered again. The students took the test in exactly the same order as before.

During the third and fourth testing periods, the repeated throws test was administered. The same organization of four squads was utilized, but there were three testing stations. The squads rotated as shown in the

following diagram:



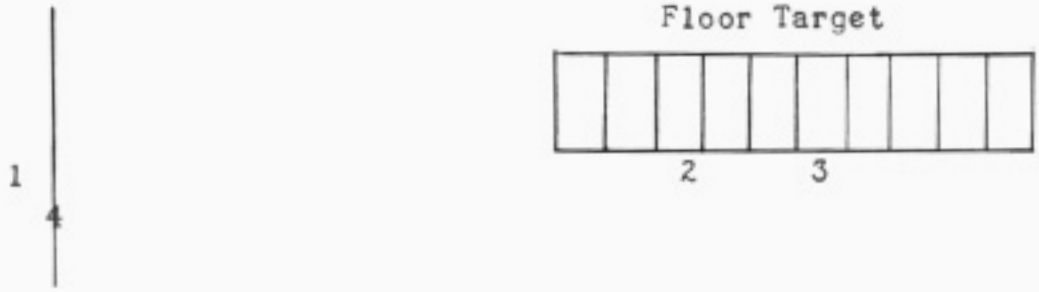
The first group tested included A (to station I), B (to station II), and D (to station III). The second group tested included C (to station I), E (to station II), and F (to station III). The same rotational procedure continued until all girls in all four squads had completed the test. This entire order of rotation was maintained for both days.

Immediately following these testing periods, the data were treated statistically. The results seemed to indicate that additional trials should be given for the repeated throws test and for the fielding test.

For this reason, the above two tests were repeated during the following two class periods. Two squads took the fielding test and then the repeated throws test, and the other two squads took the tests in the reverse order. The same order was maintained for both days.

When weather permitted meeting the class outdoors, the baserunning test was given. The same organization of four squads was utilized.

The students were trained to score the tests and did all scoring for the repeated throws test, the fielding test and the target test. For the repeated throws test, as one girl took the test, a second counted the total number of hits; a third counted the errors as listed in the directions for the test; and a fourth acted as recorder. The writer acted as timer. The same procedures were followed for the fielding test. The following diagram shows how the scorers were stationed for the target test:



Girl number one called all scores of hits on the wall. Girls numbered two and three watched the target on the floor while girl number four watched the testee to see that she did not step on or over the line. Girl number four also recorded all scores.

The writer did all timing for the baserunning test. The time was given to a student who acted as recorder.

Statistical design. The Pearson Produce Moment coefficient of Correlation was used to compute reliabilities

between the following:

1. The scores of the first trial for the repeated throws test with the scores of the second trial for the repeated throws test.
2. The sum of the scores of the first and the third trials for the repeated throws test with the sum of the scores of the second and fourth trials for the repeated throws test (odd-even method).
3. The sum of the scores of the first and the second trials for the repeated throws test with the sum of the scores of the third and the fourth trials for the repeated throws test.
4. The scores of the first trial for the fielding test with the scores of the second trial for the fielding test.
5. The sum of the scores of the first and the third trials for the fielding test with the sum of the scores for the second and fourth trials for the fielding test (odd-even method).
6. The sum total of the first ten-trial test of the target test with the sum total of the second ten-trial test of the target test.
7. The sum total of the floor scores of the first ten-trial test of the target test with the sum total of the floor scores of the second ten-trial test

of the target test.

8. The sum total of the wall scores of the first ten-trial test of the target test with the sum total of the floor scores of the second ten-trial test of the target test.
9. The sum of the scores of the first and third trials of the baserunning test with the sum of the scores of the second and fourth trials of the baserunning test.

The Pearson Product Moment Coefficient of Correlation was used to compute validities between the following:

1. Teacher ratings of softball playing ability with the sum of the scores of the four trials for the repeated throws test.
2. Teacher ratings of softball playing ability with the sum of the scores of the four trials for the fielding test.
3. Teacher ratings of softball playing ability with the scores of the first ten-trial target test.
4. Teacher ratings of softball playing ability with the better score of the two ten-trial target tests.
5. Teacher ratings of softball playing ability with the better score of the two floor scores of the ten-trial target tests.

6. Teacher ratings of softball playing ability with the better score of the two wall scores of the ten-trial target tests.
7. Teacher ratings of softball playing ability with the best score of the four trials of the base-running test.

CHAPTER IV

PRESENTATION AND DISCUSSION OF DATA

The purpose of this study was to establish reliabilities and validities for selected softball achievement tests which could be used at the University of Minnesota. Four achievement tests were devised and administered to approximately thirty college women. The data were treated statistically using the Pearson Product Moment Coefficient of Correlation.

Reliability coefficients were computed first. If a test shows a low reliability correlation, the validity correlation will be as low or lower; thus, the test would be of no use. Clarke[1] evaluates reliability coefficients in the following way.

.95-.99	Very high; rarely found among present tests.
.90-.94	High; equaled by a few of the best tests.
.80-.89	Fairly high; fairly adequate for individual measurement.
.70-.79	Rather low; adequate for group measurement but not very satisfactory for individual measurement.

---

[1]Harrison Clarke, Application of Measurement to Health and Physical Education, (New York: Prentice-Hall, Inc., 1959), p. 36.

Below .70 . . . . . Low; entirely inadequate for individual measurement although useful for group averages and school surveys.

Table I deals with the reliabilities for all achievement tests. (See Appendixes for raw score data.) The correlation coefficient for the first two trials only was determined for the repeated throws test. As shown in the table, this correlation was .63. This is below the acceptable minimum but authorities have shown that reliability can be increased by increasing the number of trials; therefore, two more trials were administered. These results were then treated using two different groupings; the sum of trials I and II with the sum of trials III and IV and the sum of trials I and III with the sum of trials II and IV. The best correlation coefficient was found when comparing the sum of trials I and III with the sum of trials II and IV. When the Spearman-Brown Prophecy Formula was used, the test yielded a coefficient of .85, which indicates a fairly high degree of reliability.

The fielding test, like the repeated throws test, yielded a low correlation (.53) when only two trials were used. For this reason, two additional trials were administered with a resulting correlation of .75 (using the odd-even method). When the Spearman Brown Prophecy Formula was applied, the correlation was increased to .86.

TABLE I

RELIABILITY COEFFICIENTS FOR THE SOFTBALL ACHIEVEMENT TESTS ADMINISTERED TO COLLEGE WOMEN AT THE UNIVERSITY OF MINNESOTA

Test	Variables	N	r	Adj.[1]
Repeated Throws	Trial I and trial II	29	.63	.78
Repeated Throws	Sum of trials I and II and sum of Trials III and IV	28	.67	.80
Repeated Throws	Sum of trials I and III and sum of trials II and IV	28	.74	.85
Fielding	Trial I and trial II	28	.53	.69
Fielding	Sum of trials I and III and sum of trials II and IV	27	.75	.86
Target	Trial I and trial II of the wall test	28	.68	.81
Target	Trial I and trial II of the floor test	28	.84	.91
Target	Trial I and trial II of the total test	28	.70	.82
Baserunning	Sum of trials I and III and sum of trials II and IV	25	.95	---

[1]Spearman Brown Adjustment Correlations.

Since the target test has three components, three reliability correlations were ascertained: the wall test, the floor test, and the total test. The correlation for the floor was .84 (.91 when the Spearman Brown Prophecy Formula was applied). This is high, with only a few tests being in the range between .90 and .94 (see page 32). The correlation for the wall test was only .68 (.81 using the Spearman Brown Prophecy Formula). Thus, the total test correlation was lowered. Nevertheless, a total test correlation of .70 was obtained and was increased to .82 by application of the Spearman Brown Formula. This correlation ranks as fairly high by the standards set forth by Clarke.

The validity correlations for all tests were based on comparisons between the individual's test scores and teacher ratings of individual performance. Two factors were considered to determine the type of rating to be used. They were the duration of the observation of the rater and the number of raters. Each has its advantages and disadvantages. One advantage of having a number of raters is that different viewpoints will be expressed due to differences in training. Another advantage is that the raters can be stationed so that each play is seen from a variety of vantage points. A disadvantage in having a number of raters is that difficulties arise in interpreting the rating sheet. These difficulties arise even

when the rating sheet seems to be in an objective form. Other disadvantages are to find competent raters, ones who would be free at a particular hour, and ones who could give that hour over a period of five or six weeks. It is imperative for all raters to be present at the same times.

Skills in some activities can be demonstrated in a relatively short period of time. This is true for the most part of individual and dual sports since the person to be observed is involved in all or a good portion of the activity. This is not true in a sport such as softball. Opportunities to observe a particular player enough times to make an adequate evaluation of his playing ability might require many observation periods. In addition, day-to-day differences in demonstrating ability are important for an adequate evaluation.

After careful consideration of these two aspects, the writer concluded that the advantages of one rater over a long period of time outweighed the disadvantages.

These teacher ratings were done for approximately the last five weeks of the quarter. The ratings were based on observations of each girl in the game situation. Each girl was observed while playing a variety of positions and received one rating of over-all playing ability. Every opportunity to handle the ball was noted. Then, a notation was made regarding the manner in which the ball was handled. It was also noted when a girl should have handled the ball.

In addition, general position play such as covering a base at the proper time or backing up another player was considered.

The writer felt that an adequate amount of time for rating was alloted. The class was divided into three teams. Two teams opposed each other while members of the third team acted as umpires or scorers, or practiced in another area of the field. The class met three times a week. With a rotation in playing, the writer was able to observe the members of each team two times a week.

Table II lists the validity correlations for each of the four tests. The first figure reports the correlation for the repeated throws test. To obtain this correlation, the writer's ratings were compared with the sum total of the four trials of the test for each student. The comparison yielded a validity correlation of .678.

The second item in the table deals with the fielding test. As with the repeated throws test, the skill ratings were compared with the sum total of the four trials of the test for each student. A correlation of .689 was obtained.

Items three, four, five, and six on the table give correlations for the target test. Item three relates to scores for the first trial only and yielded a correlation of .542. While this is not a high correlation, a review

TABLE II

VALIDITY COEFFICIENTS FOR THE SOFTBALL ACHIEVEMENT TESTS ADMINISTERED TO COLLEGE WOMEN AT THE UNIVERSITY OF MINNESOTA

Performance Evaluation	Test Score	N	r
Teacher ratings	Repeated throws test	29	.68
Teacher ratings	Fielding test	29	.69
Teacher ratings	Trial I of target test	29	.54
Teacher ratings	Better score of both trials for target test	29	.78
Teacher ratings	Better score of both trials for floor test	29	.76
Teacher ratings	Better score of both trials for wall test	29	.74
Teacher ratings	Best score of four trials for baserunning test	29	.35

of the skills tests which are presented in a measurement and evaluation book by Scott and French[1] points out the fact that the majority of the skills which are currently accepted for use have validity correlations which range from .40 to .60, with some tests falling below .40 and as low as .30. Since this test, comparing trial I with trial II, had an acceptable reliability coefficient (.81 with the Spearman Brown adjustment), it would be possible to administer only one trial if the person giving the test would be satisfied with a lower validity correlation. If, however, a higher validity correlation were desired, two administrations of the test would be necessary. When comparing the writer's ratings with the better score of the two trials, a correlation of .77 was obtained. Since a low validity coefficient was obtained when using trial I scores only, no computation was done with the wall test and the floor test using these scores. Item five reports the correlation between the skill ratings and the better score of the two obtained for the floor test. The validity coefficient was .76. When comparing the writer's ratings with the better score of the two trials of the wall test, a correlation of .74 was obtained.

---

[1]M. Gladys Scott and Esther French, Measurement and Evaluation in Physical Education, (Dubuque, Iowa: Wm. C. Brown Company), pp. 143-235.

The last item of Table II reports the validity correlation for the baserunning test. A correlation of .35 was obtained when each student's best score was compared with the writer's ratings.

While a comparison with other skills tests in general is important, comparisons with other softball skills tests are more relevant to this paper. A review of the writer's repeated throws test and that reported by Scott and French, (see pp. 7-8) permits the following statistical comparisons. Although the dimensions for the two tests were different (see pp. 7 and 22), both yielded an identical reliability correlation of .78 when trial I was compared with trial II. When statistical treatment was applied to the over-all test and the Spearman Brown Formula was used to equalize the number of trials, the writer obtained a correlation of .89 for reliability while Scott and French report a .94 for the same procedure. A comparison of the validity correlations of the two tests reveals that the writer's test has a higher correlation (.68) than either correlation reported by Scott and French (.64 for junior high school girls and .51 for college women). The correlation for the junior high school girls compares more favorably with that of the writer's than does the correlation for college women. It is possible that the reason for this lies in the dimensions of the test; that the test described by Scott and French is

more challenging to junior high school girls and thus is more discriminating at that level. The writer's test raises the line on the wall from seven and one half feet to ten feet and increases the distance from the wall by eight feet. It is the writer's opinion that the differences in dimensions is the factor influencing the increased validity correlation.

The writer feels that no real comparison can be made between the fielding tests since they are constructed in two entirely different ways.

When comparing the writer's baserunning test with that described by Thomas (see p. 13), several differences are seen. The test described by Thomas involves swinging a bat and running around all bases; whereas, in the writer's test, the testee runs only from first base to third base. The latter yielded a much higher reliability correlation (.954 as compared to .71); however, the validity correlation was much lower (.35 as compared to .545).

Although the fielding test as described by Scott and French (see pp. 10-13), includes a target, no statistics were reported on just that portion of the test. The only target test that the writer was able to find was one administered by Broer (see p. 18). This test was to determine reliability only and used junior high school girls as subjects. The reliability correlation obtained

from this test was .42. The accuracy (wall) portion of the writer's test yielded a reliability coefficient of .81. In Broer's test, however, the students threw with an underhand motion; whereas the students taking the writer's test were permitted to throw with either an overhand or a sidearm motion. Both are tests of accuracy, but it is the writer's opinion that it is more difficult to be consistent in one's accuracy when throwing with an underhand motion than it is when throwing with an overhand or sidearm motion. Thus, this may be the factor which accounts for the difference in reliabilities (.81 for the writer's test as opposed to .42 for Broer's test). Since the other component of the writer's test was power, no further comparisons can be made.

CHAPTER V

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

Summary and conclusions. The purpose of this study was to establish reliabilities and validities for selected softball achievement tests which could be used at the University of Minnesota. Four achievement tests were devised and administered to approximately thirty college women.

The data were treated statistically using the Pearson Product Moment Coefficient of Correlation. The statistical results are summarized below.

1. Target test--the writer feels that this test, using the entire test or either of its components, is highly satisfactory for use at the University of Minnesota.
   a. The wall test, which is designed to measure accuracy yielded a reliability coefficient of .81 and a validity coefficient of .74.
   b. The floor test, which is designed to measure power or distance, yielded a reliability coefficient of .91 and a validity coefficient of .76.
   c. The total test, which includes both components, yielded a reliability coefficient of .82 and a validity coefficient of .78.

2. Fielding test--the writer feels that this test, with a reliability correlation of .86 and a validity correlation of .69, certainly meets the requirements for a skills test to be used at the University of Minnesota.

3. Repeated throws test--the writer feels that this test is satisfactory for use at the University of Minnesota. Although the reliability correlation of this test is lower than that of the test reported by Scott and French (.89 as opposed to .94), even when the trials are equalized, the new test is preferred for two reasons. One important consideration is the fact that the writer's test is geared to the facilities which are now present at the University of Minnesota. The other important factor to consider is the difference between the validity correlations. The correlation obtained by the writer is .68 as opposed to only .51 for the test reported by Scott and French.

4. Baserunning test--although the reliability correlation for this test was extremely high (.95), the writer feels that the validity coefficient (.35) is too low to be of use.

Recommendations. The writer strongly recommends that the target test, the fielding test, and the repeated throws test be put into use for non-professional softball classes at the University of Minnesota.

It is further recommended that research be done to determine reliabilities and validities of these tests for use in major or professional softball classes at the University of Minnesota.

Further research should also be conducted to develop a complete battery of softball skills tests which have an adequate individual reliability and validity correlation as well as desirable inter-correlations.

Finally, the writer believes that there is a need to develop achievement tests in physical education for skills other than softball.

BIBLIOGRAPHY

BIBLIOGRAPHY

A. BOOKS

Bovard, John, Frederick W. Cozens, and E. Patricia Hagman. Tests and Measurements in Physical Education. Philadelphia: W. B. Saunders Company, 1949.

Brownell, Clifford Lee and E. Patricia Hagman. Physical Education--Foundations and Principles. New York: McGraw-Hill Book Company, Inc., 1951.

Clarke, H. Harrison. Application of Measurement to Health and Physical Education. New York: Prentice-Hall, Inc., 1959.

Cozens, Frederick W. Problem Manual in Physical Education Measurements. Philadelphia: W. B. Saunders Company, 1932.

Glassow, Ruth B. and Marion Broer. Measuring Achievement in Physical Education. Philadelphia: W. B. Saunders Company, 1938.

Graybeal, Elizabeth. The Measurement of Outcomes of Physical Education for College Women. Minneapolis: The University of Minnesota Press, 1937.

Larson, Leonard A., and Rachael Dunaven Yocum. Measurement and Evaluation in Physical, Health, and Recreation Education. St. Louis: C. V. Mosby Company, 1951.

McCloy, Charles Harold. Tests and Measurements in Health and Physical Education. New York: F.S. Crofts and Company, 1942.

_____, and Norma Dorothy Young. Tests and Measurements in Health and Physical Education. New York: Appleton-Century-Crofts, Inc., 1954.

Mathews, Donald K. Measurement in Physical Education. Philadelphia: W. B. Saunders Company, 1958.

Scott, M. Gladys and Esther French. Better Teaching Through Testing. New York: A. S. Barnes and Company, 1945.

_____. Evaluation in Physical Education. St. Louis: The C. V. Mosby Company, 1950.

_____. Measurement and Evaluation in Physical Education. Dubuque, Iowa: Wm. C. Brown Company Publishers, 1959.

Stroup, Francis. Measurement in Physical Education. New York. The Ronald Press Company, 1957.

Willgoose, Carl. Evaluation in Health Education and Physical Education. Philadelphia: W. B. Saunders Company, 1959.

B. PERIODICALS

Broer, Marion R. "Reliability of Certain Skill Tests for Junior High School Girls," Research Quarterly, XXIX (May, 1958), 139-45.

Fox, Margaret and Olive Young. "A Test of Softball Batting Ability," Research Quarterly, XXV (March, 1954), 26-7.

Liba, Marie. "A Trend Test as a Preliminary to Reliability Estimation," Research Quarterly, XXXIII (May, 1962), 245-8.

Rodgers, Elizabeth and Marjorie Heath, "An Experiment in the Use of Knowledge and Skill Tests in Playground Baseball," Research Quarterly, II (December, 1931).

Scott, M. Gladys. "The Use of Skill Tests," Journal of Health, Physical Education, and Recreation, IX (June, 1938), 364-6.

Thomas, Jesseline. "Skill Tests," Official Softball-Volleyball Guide. New York: A.S. Barnes and Company, 1943.

APPENDIXES

APPENDIX A

RAW SCORES FOR EACH SUBJECT FOR THE BASERUNNING TEST

Subject	Trials I	II	III	IV
1.	7.25	7.25	7.50	7.60
2.	7.30	7.30	7.50	7.15
3.	6.80	7.20	7.25	7.10
4.	---	---	6.80	6.50
5.	7.80	7.80	8.10	8.10
6.	---	---	6.90	6.70
7.	7.70	7.30	8.20	8.00
8.	7.50	7.30	7.50	7.25
9.	8.40	8.60	8.50	8.70
10.	7.50	7.50	---	---
11.	7.20	7.60	7.90	7.80
12.	7.20	7.40	---	---
13.	7.70	7.90	7.10	7.00
14.	7.50	7.40	---	---
15.	8.00	7.80	7.50	7.70
16.	7.90	7.70	7.70	7.60
17.	7.40	7.30	7.40	7.10
18.	7.80	7.30	7.15	7.30
19.	7.25	7.20	7.30	7.30
20.	8.20	8.10	8.20	8.20
21.	7.40	8.10	6.80	7.10
22.	7.80	8.00	7.70	7.70
23.	7.80	7.60	7.80	7.50
24.	8.10	7.70	8.10	8.10
25.	7.60	7.70	7.40	7.70
26.	8.20	7.80	7.50	7.20
27.	7.60	7.30	7.30	7.35
28.	8.50	8.70	8.90	9.00
29.	---	---	7.30	7.00
30.	7.80	7.60	8.10	8.00
31.	7.60	7.30	7.60	7.50

APPENDIX B

RAW SCORES FOR EACH SUBJECT FOR THE FIELDING TEST

Subject	Trials I	II	III	IV
1.	15	14	13	14
2.	12	12	13	16
3.	8	11	14	14
4.	10	10	16	17
5.	10	10	14	13
6.	10	11	12	14
7.	14	9	11	12
8.	10	11	11	10
9.	12	11	13	11
10.	8	8	12	10
11.	9	13	11	14
12.	12	13	12	13
13.	11	12	14	15
14.	10	9	11	13
15.	10	12	11	13
16.	10	10	11	13
17.	10	7	10	12
18.	6	13	12	14
19.	14	12	8	11
20.	10	13	12	10
21.	9	10	8	11
22.	10	8	8	9
23.	8	10	---	---
24.	10	10	8	11
25.	8	10	12	11
26.	8	11	13	9
27.	11	10	10	9
28.	10	11	12	12
29.	11	11	11	11
30.	8	7	11	13

APPENDIX C

RAW SCORES FOR EACH SUBJECT FOR THE REPEATED THROWS TEST

Subject	Trials I	II	III	IV
1.	11	12	8	11
2.	9	12	11	11
3.	11	9	12	12
4.	9	11	13	13
5.	9	11	9	11
6.	10	10	5	9
7.	11	11	8	7
8.	7	11	10	11
9.	9	10	8	10
10.	11	9	11	10
11.	10	11	11	10
12.	10	10	12	11
13.	10	10	10	9
14.	9	10	11	11
15.	9	9	9	9
16.	10	9	11	11
17.	5	8	8	10
18.	9	9	9	10
19.	7	5	8	8
20.	9	9	9	10
21.	9	10	9	10
22.	9	9	10	10
23.	9	9	9	9
24.	3	7	6	8
25.	8	10	8	10
26.	6	8	9	10
27.	7	9	11	11
28.	6	5	5	9
29.	4	7	8	8

APPENDIX D

RAW SCORES FOR EACH SUBJECT FOR THE TARGET TEST

Subject	Trial I			Trial II		
	Wall	Floor	Total	Wall	Floor	Total
1.	35	39	75	32	38	70
2.	32	36	68	34	34	68
3.	27	42	69	15	35	50
4.	26	40	66	26	35	61
5.	24	36	60	9	26	35
6.	18	22	40	25	35	60
7.	28	21	49	35	28	63
8.	17	16	33	27	28	55
9.	32	41	73	34	41	75
10.	13	36	49	19	37	56
11.	27	30	57	28	30	58
12.	32	26	58	28	35	63
13.	20	32	52	25	36	61
14.	30	24	54	28	30	58
15.	18	11	29	27	34	61
16.	25	24	49	20	23	43
17.	10	17	27	14	17	31
18.	26	22	48	39	28	67
19.	14	14	28	12	11	23
20.	28	32	60	29	29	58
21.	25	30	55	21	26	47
22.	25	32	57	22	28	50
23.	32	28	60	27	32	59
24.	19	24	43	8	20	28
25.	26	21	47	22	25	47
26.	20	16	36	30	23	53
27.	14	28	42	11	11	22
28.	14	18	32	12	12	24
29.	15	10	25	13	8	21